Research Statement

Technology innovations allow scientists to collect massive amount of data with complex structures. The non-standard features of such data include non-normality, unbalanced designs, complex heterogeneity and dependence structures, and high-dimensionality. My work in the last a few years have focused on tackling challenging statistical problems arising from data with such non-standard features.

My research began with my work as a graduate student on nonparametric analysis of nonstandard experimental data, which in many aspects do not comply with traditional modeling assumptions. And I have since expanded my research into two new directions: model checking and model selection. Model assessment and selection is crucial for constructing valid statistical models that have the ability to identify scientifically meaningful structures in noisy, complex data. My work in these two directions emphasizes developing novel theory and methodology that can accommodate complex and realistic data structures, such as those that exhibit non-normality and outlier contamination or those that exhibit correlation among observations. Recently, quantile regression caught my interest for its powerful ability to incorporate general heterogeneity and to reveal important underlying structures in the data that would otherwise go undetected.

My research is currently supported by an NSF grant for which I am serving as the principle investigator. The statistical methods I have developed have applications across many disciplines, including biomedical studies, sociology, economics, environmental studies, bioinformatics, among others. In what follows, I outline my research contributions and my perspectives on future work.

1. Lack-of-fit tests for model checking

Representative papers:

- Wang, L. and Qu, A. (2008) Consistent model selection and data-driven model checking for longitudinal data based on estimating equations. To appear in *Journal of the Royal Statistical Society, Series B.*
- Wang, L. and Qu, A. (2007) Robust tests in regression models with omnibus alternatives and bounded influence. *Journal of the American Statistical Association*, 102, 347-358.
- Wang, L. and Zhou, X. H. (2007) A smoothing-based nonparametric test for diagnosing the variance structure in regression models. *Biometrics*, 63(4), 1218-1225.

- Wang, L. and Van Keilegom, I. (2007) Nonparametric trend detection with time series errors. *Statistica Sinica*, 17, 369-386.

Regression analysis is fundamental for statistical analysis. The importance of testing the validity of a specified regression model can not be overstated. A misspecified model often leads to biased estimation and misleading inferences. Classical lack-of-fit tests assume a parametric alternative. They are powerful when the parametric alternative is correctly specified but may lose power considerably when the alternative is misspecified. My research in this area mainly focuses on developing modern nonparametric lack-of-fit tests that are consistent against flexible alternatives.

Although such nonparametric tests have been studied for iid data, many important problems remain unsolved for more complex data structures. One such challenging problem arises in the area of correlated data, where only sparse work is available for model checking. Correlated data frequently occur in many fields. For example, in the large scale Wisconsin Epidemiologic Study of Diabetic Retinopathy, both eyes of each of the 720 individuals in the study were examined for the presence of diabetic retinopathy. The responses from the two eyes of the same individual are correlated. In general, dependence structure arises naturally when repeated measurements are taken on the same subjects over time or when observations are made on each individual within a cluster. The main challenge for developing model checking procedures for correlated data is that it is often difficult or impossible to write down the likelihood function, especially if the response variable is discrete. Motivated by recent development in quadratic inference function analysis of correlated data, Wang and Qu (2008) proposed a novel data-driven lack-of-fit test. The new test displays high power against various types of alternatives.

In practical applications, it is common to observe data that are contaminated by outliers. Standard model checking procedures are rather sensitive to even a small percentage of contamination. Their performances sometimes totally break down. Wang and Qu (2007) proposed a new class of smoothing-based nonparametric lack-of-fit tests, which enjoys the nice omnibus property and at the same time possesses desirable robustness features. To the best of my knowledge, this is the first work in the literature that bridges robust tests with modern smoothing-based omnibus tests.

In addition to the above, I have also worked on model checking for heteroscedastic data (Wang and Zhou, 2007) and model checking for regression models with an autoregressive error structure (Wang and van Keilegom, 2007).

2. Model selection

Representative papers:

- Wang, L. and Li, R. (2008) Weighted Wilcoxon-type smoothly clipped absolute deviation method. To appear in *Biometrics*.
- Wang, L. (2008) Wilcoxon-type generalized Bayesian information criterion. Accepted by *Biometrika*.
- Wang, L. and Qu, A. (2008) Consistent model selection and data-driven model checking for longitudinal data based on estimating equations. To appear in *Journal of the Royal Statistical Society, Series B*.

Variable selection is an important step in model building. The area of model selection has become the focus of much research in recent years. Selection of important factors can provide a better understanding of the underlying data generating process, and improve the accuracy of prediction and estimation.

Schwartz's BIC and its variants enjoy great popularity in model selection. However, Schwartz's BIC requires an unambiguous specification of a parametric distribution and its consistency breaks down when the underlying distribution is misspecified. This motivates me to propose (Wang, 2008) a generalized Bayesian information criterion which relaxes the strong distribution assumption required by Schwarz's BIC. It also outperforms Schwarz's BIC with heavier-tailed data in the sense that asymptotically it can yield substantially smaller L_2 risk. Moreover, it can be conveniently implemented via existing statistical programs.

In Wang and Li (2008), we introduced a weighted Wilcoxon-type shrinkage procedure, which deals with robust variable selection and robust estimation simultaneously. The new procedure are effective in handling outliers in both x and y directions. Such a procedure appears to be novel for shrinkage type variable selection methods including SCAD and LASSO. Wang and Qu (2007) also successfully tackled the challenging problem of model selection for correlated data. We proposed a novel BIC-type model selection criterion, which does not require the full likelihood or quasilikelihood. With probability approaching one, it selects the most parsimonious correct model. Although a working correlation matrix is assumed, there is no need to estimate the nuisance parameters in the working correlation matrix; moreover, the model selection procedure is robust against the misspecification of the working correlation matrix.

3. Analysis of nonstandard experimental data

Representative papers:

- Wang, L. and Akritas, M. G. (2006a) Testing for covariate effects in fully nonparametric ANCOVA model. *Journal of the American Statistical Association*, 101, 722-736.

- Wang, L. and Akritas, M. G. (2006b) Two-way heteroscedastic ANOVA with large number of levels. *Statistica Sinica*, 16, 1387-1408.

I began working in this area as a graduate student. My research was motivated by the observation that experimental data in many scientific fields often exhibit one or more nonstandard features such as non-normality, nonlinearity, non-parallelism and heteroscedasticity.

In Wang and Akritas (2006a), we extended the state of art of nonparametric analysis of covariance in a number of ways. The new methodology applies to testing all common hypotheses; it allows for both continuous and discrete ordinal response variable; it permits heteroscedastic and non-normal errors; and it does not require the covariate to be equally spaced or to have the same distribution at different treatment level combinations. In Wang and Akritas (2006b), we investigated the "large p, small n" problem in two-way ANOVA setting, where one factor or both factors have large number of levels. The results in this paper can be applied in many disciplines. For example, in agricultural trials it is not uncommon to see large number of treatments (cultivars, pesticides, fertilizers, etc) but limited replications per treatment combination. In a recent statewide agricultural study performed by Washington State University, 40 different varieties/lines of winter wheat are investigated.

4. Quantile regression

Representative papers:

- Wang, L. (2008) Nonparametric test for checking lack-of-fit of quantile regression model under random censoring. *Canadian Journal of Statistics*, 36(2), 321-336.

- Wang, HX and Wang, L. (2008) Locally weighted censored quantile regression. Submitted to JASA.

- van Keilegom, I and Wang, L. Semiparametric modelling and estimation of the dispersion function in regression. To be submitted soon.

Quantile regression has emerged as a significant extension of the classical least-squares regression. By estimating a family of conditional quantiles, we are capable to provide a more complete picture of the covariate effects. Quantile regression has broad applications. For instance, it has been widely used in economics to study consumer demand, determinants of wags, discrimination effects and trends in income inequality, among others.

For survival analysis, censored quantile regression offers a valuable supplement to Cox proportional hazards model. I first started to work in this area when I found that there was no literature available on checking the adequacy of quantile regression function under censoring. This led to Wang (2008), in which I derived a useful kernel-type conditional moment test. In the recent work of Wang and Wang (2008), we proposed a novel locally weighted censored quantile regression approach, which relaxes the stringent model assumptions required by many existing procedures, such as unconditional independence or global linearity. The new approach adopts the redistribution-of-mass idea and employs a local reweighting scheme. In my joint work with van Keilegom (2008), we proposed a flexible semiparametric framework for modeling heteroscedasticity, which includes in particular semiparametric quantile regression models. We obtained general asymptotic results that apply to many commonly used semiparametric structures, for instance, the partially linear structure and single single-index structure.

5. Perspectives on future research

I expect myself to continue my efforts in advancing statistical theory and methods for analyzing data with highly nonstandard features. Despite recent development, a plethora of important problems remain unsolved. Some of the problems I'd like to explore in the near future include: (1) model checking and model selection for high-dimensional correlated data; (2) robust variable selection for longitudinal data using rank-based estimating equations; (3) efficient and robust inference for varying coefficient models based on local ranks; (4) efficient estimation in heteroscedastic semiparametric quantile regression models. These problems are theoretically difficult but results from these studies will greatly enhance the tools available to scientists for analyzing complex data.

Innovations in many scientific fields constantly pose interesting new theoretical problems for statisticians; on the other hand, statisticians have great potential to apply their knowledge and help solve real problems in diverse disciplines. I have no doubt that my working in the area of statistics in the coming years will be a very rich and rewarding experience.